Species Mention Entity Recognition, Linking and Classification Using RoBERTa in Combination with Spanish Medical Embeddings

Rodrigo del Moral[1], Javier Reyes-Aguillón[1], Orlando Ramos-Flores[2], Helena Gómez-Adorno[2] and Gemma Bel-Enguix[3]

[1]*Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México*
[2]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México*
[3]*Instituto de Ingeniería, Universidad Nacional Autónoma de México*

Abstract

This paper introduces the solution submitted by puma's team to the LivingNER 2022 shared task. The proposed system finds species mentions inside clinical case reports, links the mentions to the NCBI taxonomy database, and classifies the identified species according to clinically relevant categories. Our architecture combines a Spanish biomedical RoBERTa model with a neural network layer for sequence classification. We used the NCBI taxonomy entries for the entity linking problem and calculated the Levenshtein distance to each identified species. We used the medical Spanish word embeddings to train a logistic regression classifier to identify clinically relevant categories of the species. The obtained results ranked above the mean for all tasks, confirming that a RoBERTa-based system, pre-trained with texts related to the biomedical field, performs well in the named entity recognition task. Using a dictionary of terms in the entity linking task proved to improve the performance.

Keywords

LivingNER, Named Entity Recognition, Entity Linking, Machine Learning, RoBERTa

1. Introduction

Named entity recognition within medical texts, from both scientific and clinical fields of expertise, is a highly challenging task. Clinical texts are diverse in language and details, containing ad hoc terminology, acronyms, and jargon. They are often written in a hurry because it consumes time away from the patient's care. Several shared tasks have been organized to motivate the research on named entity recognition (NER) over clinical texts. For example, CLEF eHealth 2021 [1] promoted NER and entity classification in Spanish language texts belonging to the area of radiology. IberLEF eHealth-KD 2019 [2], 2020 [3], and 2021 [4] focused on the identification and classification of entities within articles extracted from the PubMED library, as well as from WikiNews and the CORD-19 corpus of COVID-19-related scientific resources. Entity recognition

IberLEF 2022, September 2022, A Coruña, Spain.
rodrigodelmoral@comunidad.unam.mx (R. d. Moral); jav_15@comunidad.unam.mx (J. Reyes-Aguillón); orlando.ramos@aries.iimas.unam.mx (O. Ramos-Flores); helena.gomez@iimas.unam.mx (H. Gómez-Adorno); gbele@iingen.unam.mx (G. Bel-Enguix)
0000-0003-1868-9013 (R. d. Moral); 0000-0002-6205-2610 (J. Reyes-Aguillón); 0000-0002-8579-4123 (O. Ramos-Flores); 0000-0002-6966-9912 (H. Gómez-Adorno); 0000-0002-1411-5736 (G. Bel-Enguix)

has also been applied in other health-related fields, such as CLEF 2020's CheMU shared task [5], which sought to identify entities related to chemical reactions.

This increased attention to the subject comes with a growth of biomedical corpora available on the web and an understanding of the benefits that organizing information from all medical texts can bring to research. Awareness of the advantages of linking concepts from disorganized texts to knowledge bases is essential. Firstly, we could delve deeper into such concepts and access structured data from other related bases, which is very valuable for multidisciplinary research. Another benefit of medical text analysis systems is extracting information in large volumes. With enough data to train machine learning algorithms, there is a possibility to generate knowledge in a semi-supervised or unsupervised form.

This paper introduces a solution for identifying species mentioned in clinical texts. Our system comprises a Spanish language RoBERTa model and a neural network layer trained for token classification. Moreover, our system links the identified species to the NCBI taxonomy database using a dictionary of ID codes. Finally, our system classifies the found species into four relevant categories using a Logistic Regression Classifier and a Word Embedding model. A more detailed explanation of the LivingNER shared task, as well as its data set and results, can be found in the overview paper [6].

The remainder of this article is structured as follows. In Section 2 we describe the data set used for the task and briefly discuss its creation. In Section 3 we describe the models proposed for the Named Entity Recognition, Entity Linking, and Entity Classification subtasks, as well as pre-processing and post-processing of the data. In section 4, we describe the experiments carried out with the models and report their results. Finally, in Section 5 we discuss the obtained results and give ideas for future system improvement.

2. Data Set

The LivingNER corpus [7] is composed of 2,000 clinical cases from 20 medical specialties annotated for species (living organisms and microorganisms) and infectious diseases mentions. All texts are in Spanish, and the annotations were manually generated by a specialist physician and reviewed by another clinical specialist. The corpus creation process was carried out in five months with the use of the *brat* rapid annotation tool [8]. The specialists mapped the annotated entities to the NCBI (National Center for Biotechnology Information) taxonomy database. Finally, half of the documents had their entities classified into four categories relevant to the clinical field: pets and farm animals, animals that can cause injuries, food, and nosocomial infections. The final version of this corpus contains 30,886 species mentions (43.9% of which are human mentions) and 11,841 infectious diseases mentions. Of these, 31,694 are mapped to the NCBI taxonomy database, and 8,673 are unique mentions.

The entire corpus was divided into three parts (training, validation, and testing) to leave data out to evaluate the shared task. The training set consists of 1,000 annotated documents given to the participants to train the models. The validation set consists of 500 annotated documents. Finally, the participants were given 490 unannotated documents to evaluate the models. However, only half of the documents had their entities classified into medically relevant categories. In addition, the organizers incorporated 12,982 background documents into the test

set. The participants had to process the 13,472 documents. However, the background data was not considered for the model evaluations.

3. Methodology

The LivingNER task is itself composed of three smaller subtasks. We designed and implemented a Named Entity Recognition system for the first subtask that identifies species in the documents provided in the LivingNER corpus. We propose an Entity Linking system based on a species-ID dictionary and a Levenshtein distance search for the second subtask. For the third subtask, we conceived a Logistic Regression Classifier trained with vectors from a Word Embedding model trained on medical texts.

3.1. Subtask 1: Named Entity Recognition

The system designed for the NER subtask is composed of three stages. First, pre-processing converts the raw documents and provides annotations to a format suitable for the main model. Then, the token classification model is trained using data from the LivingNER training set, and the testing data set is processed through the resulting model. Finally, the post-processing stage prepares a file compliant with the format required by the evaluation script. Figure 1 shows the steps for the entire process.

3.1.1. Pre-processing

The annotated documents in the LivingNER corpus contain species mentions in the *brat standoff* format, with each entity referenced to the document where it originates. We mapped the annotations to an *IOB* representation to produce a suitable input for the entity recognition model. The process consisted of tokenizing all entities for the program to distinguish between beginning-tokens and inside-tokens (in the case of multi-word entities). Subsequently, the annotated entities were aligned with their positions in the original texts using the offsets provided by the *brat* format. Finally, we tokenized the portions of the documents with no labels, and the tokens were assigned the outside-token IOB-tag.

The tokenizer employed ignores periods and line breaks while maintaining other non-alphanumeric characters such as hyphens, parentheses, and commas. Therefore, these symbols were treated as single tokens and annotated with the inside-token IOB-tag. The decision to keep such symbols was taken since some of these characters appear to provide valuable contextual information for the model.

Finally, all documents were divided into sentences to perform the model training and entity prediction task more efficiently.

3.1.2. RoBERTa-based Classifier

Once the clinical texts were split and mapped to the IOB format, we carried out a series of tests to pick the base algorithm for the NER system. A more detailed description of these tests can be found in Section 4.

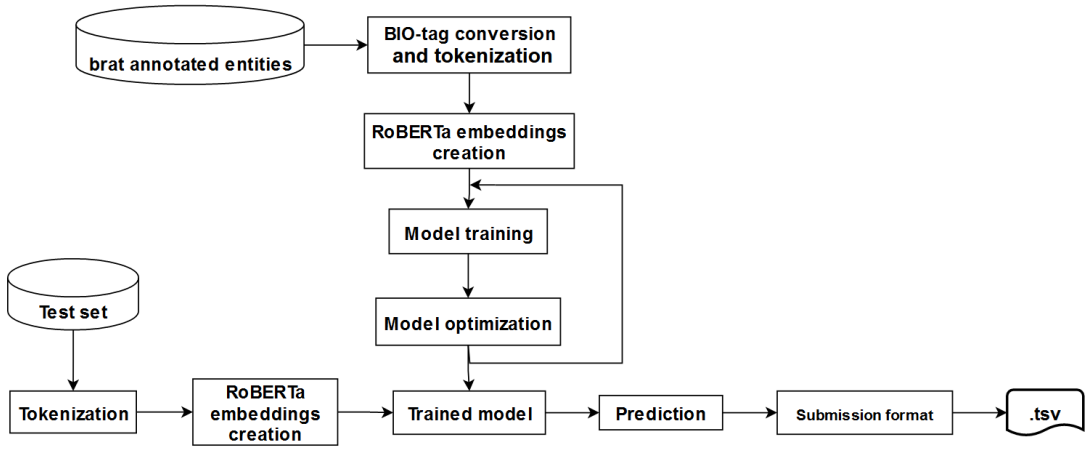

Figure 1: Named Entity Recognition System Diagram (Subtask 1)

The Named Entity Recognition subtask aims at identifying all species mentioned (human and non-human) inside a set of clinical report documents. An IOB tagging scheme was chosen to separate the tokens into five different classes. The "B-HUM" and "I-HUM" tags are used if the token belongs to a human mention, where the former is used when the token is the beginning (or only word) of the mention and the latter when it is an inside token. The "B-SPC" and "I-SPC" tags are used similarly for the non-human species mentioned. And finally, an "O" tag is used when the token does not belong to any species mention.

Considering the experiment results, we chose a RoBERTa model as the basis of our system. The particular chosen model [9] is pre-trained with biomedical texts written in Spanish. On top of this RoBERTa model, a token classification layer was trained with the IOB tags from the LivingNER training annotations.

3.1.3. Post-processing

Once predicted, the output of the RoBERTa algorithm underwent a thorough post-processing stage. This post-processing was necessary to convert the predicted IOB tags to a submission-ready file that follows the *brat standoff* format.

First, we joined the sub-embeddings generated by the RoBERTa tokenizer, taking into account only the predicted label for the first sub-embedding and extending it through all the other fragments. Then, we applied a series of rule-based filters, discarding every token that presented the "O" label. We also discarded labels that were predicted as inside-tokens despite not being immediately preceded by beginning-token or inside-token labels. Finally, we dropped entities that consisted entirely of stop-words, numbers, and punctuation symbols.

Further on, we joined the multi-word predicted entities and discarded their IOB tags. After that, we aligned the predicted entities with the original texts.

Finally, all the required characteristics of the found entities were gathered in a *.tsv document for submission and evaluation.

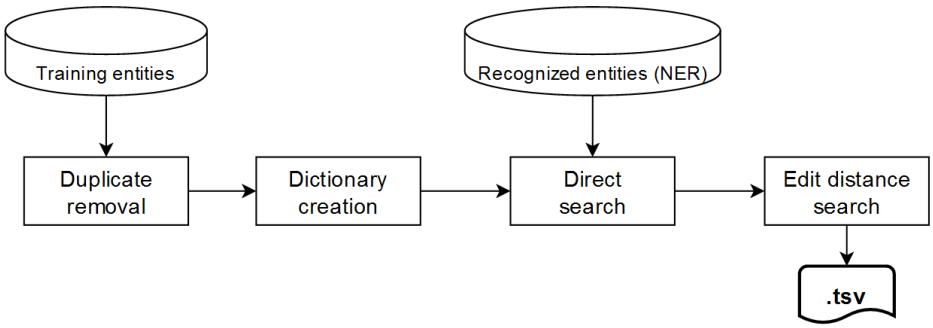

Figure 2: Entity Linking System Diagram (Subtask 2)

3.2. Subtask 2: Entity Linking

The proposed system for the Entity Linking subtask also comprises three stages. First, a pre-processing stage converts the output from the first subtask into a list of humans and species names. Then, the codes of each entity are searched through the IDs dictionary. Finally, the post-processing stage prepares the submission file with the predicted codes included. Figure 2 depicts the steps for the entire process.

3.2.1. Pre-processing

The entities in the LivingNER training set were extracted to form a dictionary of known entities. The entities in this dictionary are already linked with their respective NCBI taxonomy IDs. The process then drops the capitalization from the identified species and removes all the duplicated keys.

3.2.2. Entity Linking

As a first step, each entity's span was converted to lowercase and searched inside the dictionary of known NCBI IDs. The system then assigned the same code and properties to the entities that produced an exact match with the species in the dictionary.

As a second step, we implemented a Levenshtein distance lookup for all of the entities that did not produce an exact match. This lookup algorithm compares every remaining entity against all the keys from the dictionary of known entities. Finally, we assigned a taxonomy ID depending on a threshold distance.

3.2.3. Post-processing

The post-processing step for the entity linking model consisted of creating a new *.tsv file. In this file, we appended the data from the NER sub-task with the NCBI code predictions.

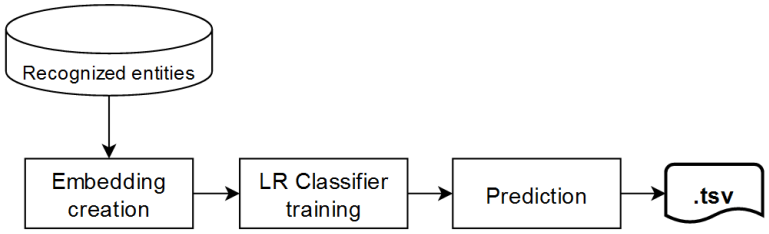

Figure 3: Entity Classification System Diagram (Subtask 3)

3.3. Subtask 3: Entity Classification

For the Entity Classification subtask, we propose a three-stage system. First, the pre-processing stage extracts the NCBI codes and species names from the Entity Linking system output. Then, the species names are mapped to a Word Embedding representation. We train a Logistic Regression Classifier in the classifying stage using the vector representations and the categories from the training data set. The testing data is then processed with this classifier. Finally, the post-processing stage prepares the submission file, including the NCBI codes found in each category. Figure 3 shows the specific steps for the entire process.

3.3.1. Pre-processing

We read the entities generated by the system from the previous sub-task in the *brat* format as plain text. Then, we tokenized each of these entities by word, discarding punctuation marks, capitalization of words, and stop-words.

Subsequently, we obtained a word embedding model pre-trained with Spanish language medical texts (SciELO-CBOW) [10] and loaded it. We used this model to map the pre-processed tokens from each entity to their vector representations. In the case of multi-word entities, we calculated an average vector from the individual tokens. The pre-trained embedding model retains the n-gram vectors. Thus, the algorithm calculated an average vector from the word n-grams in case none of the looked-up words were present in the vocabulary. Finally, we mapped the entire LivingNER training set to its embedding representation using the same algorithm.

3.3.2. Logistic Regression Classifier

We trained a binary Logistic Regression classifier for each category using the vectors retrieved from the training set. The categories of interest for the classification subtask were: *isPet* for pets and farm animals, *isAnimalInjury* for species of animals that caused injuries in a specific document, *isFood* for species that are consumed by humans (excluding medicines), and *isNosocomial* for healthcare-associated infections.

Finally, we processed the testing set, making binary predictions for the different clinically relevant categories using the embedding representations.

3.3.3. Post-processing

The post-processing results from the Logistic Regression classifier consisted of creating a new *.tsv file. This submission-ready file specifies whether each document contains species from every category and their respective NCBI taxonomy IDs.

4. Experiments and Results

We performed a series of experiments aimed at improving the model's performance. We trained three different transformer models for the validation stage for the NER subtask. Regarding the Entity Linking subtask, we only experimented with a Word Embeddings model pre-trained for medical texts. Finally, we performed a hyperparameter optimization over the Logistic Regression Classifier for the Entity Classification subtask.

4.1. Subtask 1: Named Entity Recognition

We tested the tagged documents from the pre-processing with a series of token classification algorithms. These tests were carried out to determine the optimal base model to be used by our system. Subsequently, we chose three different Transformer-based models to try out, considering state of the art in NER problem-solving. The first two models are based on mBERT [11] and BETO [12], with an additional layer for token classification using CoNLL 2002 data. While for the third model, we chose a RoBERTa-based model pre-trained with biomedical texts written in Spanish [9]. In all the models, we added an extra layer for token classification. We trained this layer with the IOB tags generated from the LivingNER training set. We used a stochastic gradient descent method with weighted decay for model optimization. Table 1 presents the parameters we used to train all the models.

Table 1
Named Entity Recognition System Training Parameters

Parameter	Value
Sentence maximum length	400
Batch size	2
Epochs	1
Learning rate	constant

Using the gold standard from the LivingNER validation set and the provided evaluation library, we evaluated the results from the three systems. These results are presented in Table 2 and include each system's micro-averaged precision, micro-averaged recall, and micro-averaged F1 score. Ultimately, we chose the RoBERTa-based model to solve the NER challenge, following the results of the experiments.

Once the RoBERTa base model was validated and selected, we processed the LivingNER test set using this model to obtain the first predictions run. Then, using the same RoBERTa pre-trained model, we trained a second system combining the training and validation sets into one unique training set. The training parameters for this second model remained the same

Table 2
Named Entity Recognition Base Models Evaluation Metrics

	Base model	MiP	MiR	MiF
1	mBERT	0.9062	0.7577	0.8253
2	BETO	0.9068	0.7843	0.8411
3	RoBERTa	0.8744	0.8991	0.8866

except for the training epochs, which we increased to 2. In Table 3, we present the evaluation metrics for the submitted results. The table is divided into results from a complete evaluation, an evaluation only for the species mentions, and an evaluation only for the human mentions. We include the same metrics as Table 2 for the NER subtask. The LivingNER organizers calculated these metrics. The table includes the mean and standard deviation of the submissions from all teams.

Table 3
Named Entity Recognition System Evaluation Metrics

		All			Species			Human		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.8714	0.8953	0.8832	0.8202	0.8502	0.8349	0.9430	0.9562	0.9496
2	RoBERTa	0.9284	0.8899	0.9087	0.9038	0.8417	0.8716	0.9598	0.9549	0.9574
	MEAN	0.8763	0.8077	0.8239	0.8112	0.7579	0.7781	0.9312	0.8750	0.8849
	STD	0.1542	0.2465	0.2371	0.2490	0.2565	0.2510	0.1156	0.2388	0.2230

4.2. Subtask 2: Entity Linking

After retrieving the species names and codes from the training and validation sets, we obtained a dictionary of codes consisting of 3,412 unique entities and NCBI IDs. Then, we pre-processed the results from the NER RoBERTa model that was trained with one epoch and the training set only. These results contained 115,120 found entities. From these entities, 80,424 were marked as human and linked directly to the respective NCBI code. Afterward, we looked up the remaining 34,696 entities marked as species on the dictionary for exact matches. In case there was no exact match for an entity, we performed a Levenshtein distance search. These two types of search combined resulted in 25,967 species found.

We only ran the entity linking system for time constraints on the first NER results set. We show the evaluation metrics for this sub-task in Table 4, including the mean and standard deviation of the submissions from all teams. This table is organized similarly to Table 3.

4.3. Subtask 3: Entity Classification

The results for the entities classification are shown in Tables 5 and 6. The Table 5 shows the evaluation taking into account the specific NCBI taxonomy IDs for each classification. While Table 6 only considers the classifications for each document.

Table 4
Entity Linking System Evaluation Metrics

		All			Species			Human		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.9389	0.8075	0.8682	0.9350	0.6972	0.7988	0.9430	0.9562	0.9496
	MEAN	0.8488	0.8068	0.8267	0.7598	0.6917	0.7228	0.9588	0.9621	0.9604
	STD	0.1577	0.1476	0.1508	0.2893	0.2582	0.2704	0.0193	0.0140	0.0151

Table 5
Entity Classification System Evaluation Metrics
(Taking into account the correctness of the linked NCBI codes)

		isPet			isAnimalInjury		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.0240	0.2500	0.0438	0.0000	0.0000	0.0000
	MEAN	0.0093	0.1023	0.0170	0.0001	0.0208	0.0002
	STD	0.0146	0.1625	0.0268	0.0002	0.0510	0.0005
		isFood			isNosocomial		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.0211	0.2692	0.0391	0.0000	0.0000	0.0000
	MEAN	0.0088	0.1538	0.0165	0.0000	0.0000	0.0000
	STD	0.0097	0.1519	0.0181	0.0000	0.0000	0.0000

Table 6
Entity Classification System Evaluation Metrics
(Taking into account only the binary classifications)

		isPet			isAnimalInjury		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.0240	0.2500	0.0438	0.0167	0.2500	0.0312
	MEAN	0.0201	0.2917	0.0369	0.0093	0.3333	0.0177
	STD	0.0166	0.3107	0.0303	0.0110	0.2458	0.0206
		isFood			isNosocomial		
	Base Model	MiP	MiR	MiF	MiP	MiR	MiF
1	RoBERTa	0.0211	0.2692	0.0391	0.0000	0.0000	0.0000
	MEAN	0.0180	0.3910	0.0340	0.0013	0.2083	0.0025
	STD	0.0178	0.4044	0.0338	0.0024	0.3323	0.0047

5. Conclusions

The proposed system uses a RoBERTa-based model pre-trained with biomedical texts written in Spanish for Named Entity Recognition. This model scored best compared to the mBERT and BETO pre-trained models we also experimented with. The species mentions recognized with

the RoBERTa model were linked to the NCBI taxonomy database using a dictionary of ID codes and the Levenshtein distance. Finally, an entity classification was developed for four entity classes, as shown in Tables 5 and 6.

The Named Entity Recognition system results scored higher than the averaged metrics. This shows that models pre-trained with documents from a specific area perform better than more general models. Still, we believe that the performance could be increased further by narrowing down the selection of medical specialties covered by the clinical notes. In the entity linking task, our team's scores also look favorable. Our Entity Linking system was very fast due to the dictionary structure and all the human entities that link directly to only one code. However, the prediction of new entities fails when they are not present in the dictionary. Regarding the last task, there is still a good opportunity for improvement. Nevertheless, we can still confirm that a simple logistic regression model cannot classify the entities correctly.

In future work, we expect to improve our Entity Linking system by applying deep learning techniques such as graph neural networks. We intend to use a transformers-based model to enhance the classification for the classification task.

Acknowledgments

This work has been carried out with the support of CONACyT-Mexico projects CB A1-S-27780, SECTEI (Mexican Government) project SECTEI/202/2021, DGAPA-UNAM PAPIIT project numbers TA400121 and TA101722, and CONACYT No.CVU. 1148113 scholarship. The authors also thank CONACYT for the computing resources provided through the Deep Learning Platform for Language Technologies of the INAOE Supercomputing Laboratory.

References

[1] V. Cotik, L. Alonso Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. Defrancesca, A. Dellanzo, M. Fernández Urquiza, Overview of CLEF eHealth Task 1-SpRadIE: A challenge on information extraction from Spanish radiology reports, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. Bucharest, Romania, September 21st to 24th, 2021., 2021, pp. 732–750. URL: http://ceur-ws.org/Vol-2936/paper-61.pdf.

[2] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, 2019, pp. 1–16. URL: http://ceur-ws.org/Vol-2421/eHealth-KD_overview.pdf.

[3] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Diaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing

(SEPLN 2020). Málaga, Spain, September 23th, 2020, 2020, pp. 1–14. URL: http://ceur-ws.org/Vol-2664/eHealth-KD_overview.pdf.

[4] A. Piad-Morffis, S. Estevez-Velarde, Y. Gutierrez, Y. Almeida-Cruz, A. Montoyo, R. Muñoz, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021) 233–242. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392.

[5] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, A. Albahem, J. Wang, Y. Ren, Z. Zhang, Y. Zhang, M. H. Dao, P. Ruas, A. Lamurias, F. M. Couto, D. Lowe, J. Mayfield, A. Köksal, H. Dönmez, O. Arzucan, D. Mahendran, G. Gurdin, N. Lewinski, C. Tang, B. T. McInnes, P. R. Rao, S. L. Devi, L. Cavedon, T. Cohn, T. Baldwin, K. Verspoor, An extended overview of the clef 2020 chemu lab: Information extraction of chemical reactions from patents, Julien Knafou 10 (2020) 11.

[6] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources, Procesamiento del Lenguaje Natural (2022).

[7] A. Miranda-Escalada, E. Farré-Maduell, G. González Gacio, M. Krallinger, LivingNER corpus: Named entity recognition, normalization & classification of species, pathogens and food, 2022. URL: https://doi.org/10.5281/zenodo.6642852. doi:10.5281/zenodo.6642852, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[8] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.

[9] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.

[10] F. Soares, M. Villegas, A. Gonzalez-Agirre, J. Armengol-Estapé, S. Barzegar, M. Krallinger, Fasttext spanish medical embeddings, 2020. URL: https://doi.org/10.5281/zenodo.3744326. doi:10.5281/zenodo.3744326, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) and the ICTUSnet project (https://ictusnet-sudoe.eu/en/).

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020, pp. 1–10.